CARF Working Paper

CARF-F-498

Epistemological Mechanism Design

Hitoshi Matsushima
University of Tokyo

Shunya Noda
University of British Columbia

First Version: October 26, 2020
This Version: February 4, 2021

CARF is presently supported by The Dai-ichi Life Insurance Company, Limited, Nomura Holdings, Inc., Sumitomo Mitsui Banking Corporation, Mizuho Financial Group, Inc., MUFG Bank, Ltd., The Norinchukin Bank and The University of Tokyo Edge Capital Partners Co., Ltd. This financial support enables us to issue CARF Working Papers.

Epistemological Mechanism Design*

Hitoshi Matsushima[†] Shunya Noda[‡]

First Version: October 26, 2020 This Version: February 4, 2021

Abstract

This study demonstrates a novel epistemological approach to mechanism design. We consider a type space in which agents are either selfish or honest, and show that a slight possibility of honesty in higher-order beliefs motivates all selfish agents to behave sincerely. Specifically, in our model, a central planner attempts to elicit correct information through mutual monitoring. We assume severe restrictions on incentive device availability: neither public monitoring nor allocation rules are available. Thus, the central planner uses only monetary payment rules. It is well-known that if "all agents are selfish" is common knowledge, eliciting correct information as unique equilibrium behavior is impossible. Nevertheless, we show a very permissive result: the central planner can elicit correct information from all agents as unique Bayes Nash equilibrium behavior if "all agents are selfish" is *not* common knowledge. This result holds even if honest agents are mostly motivated by monetary interests.

Keywords: epistemological mechanism design, unique information elicitation, common knowledge of all agents' selfishness, intrinsic preference for honesty, quadratic scoring rule.
JEL Codes: C72, D71, D78, H41.

*The first version of this study was entitled "Unique Information Elicitation," CARF-F-496, University of Tokyo, 2020. This paper extends the theory part of Matsushima and Noda (2020). This study was supported by a grant-in-aid for scientific research (KAKENHI 20H00070) from the Japan Society for the Promotion of Science (JSPS), the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of the Japanese government, and the Social Sciences and Humanities Research Council of Canada.

[†]Department of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. E-mail: hitoshi [at] e.u-tokyo.ac.jp.

[‡]Vancouver School of Economics, University of British Columbia, 6000 Iona Dr, Vancouver, BC V6T 1L4, Canada. E-mail: shunya.noda [at] gmail.com.

1 Introduction

This study demonstrates a new approach to mechanism design from an epistemological perspective. We introduce an epistemological type space in which agents are either selfish or honest (nonselfish) in higher-order beliefs. We show that a slight possibility of an honest agent in higher-order beliefs incentivizes all selfish agents to behave sincerely.

We assume that "all agents are selfish," (i.e., all agents are motivated only by their monetary interests) is not common knowledge. That is, we consider an epistemological possibility that some agents are motivated not only by their monetary interests but also by ethical or behavioral motives, such as an intrinsic preference for honesty. Many previous studies in the mechanism design literature ignored these motives. As a departure from previous literature, this study demonstrates a mechanism design method that helps elicit a nonselfish motive hidden in an agent's mind and harness it.

Specifically, we consider a problem in which a central planner attempts to elicit correct information from agents. The central planner needs to know which state of the world actually occurs, whereas there exist agents who are fully informed of it. Hence, the central planner attempts to design a mechanism to incentivize this agent to truthfully announce the state. However, we assume severe restrictions on incentive devices: no public monitoring technology is available,[1] and the central planner cannot use any allocation device besides monetary transfers.[2] The central planner in this study is permitted only to use a message-contingent payment rule.

To overcome the difficulty resulting from these restrictions, the central planner listens to the messages from multiple agents who have the same information and have them mutually monitor each other. However, for such mutual monitoring to function, the central planner still needs to overcome another challenge in incentives, that is, the multiplicity of equilibria resulting from coordination failure. Hence, this study analyzes the possibility of *unique information elicitation*, implying that the central planner elicits correct information through agents' unique equilibrium behavior.

The mechanism design literature has traditionally assumed that "all agents are selfish" is common knowledge. This assumption makes severe multiplicity of equilibria to be inevitable in our problem because agents' preferences for monetary transfers are independent of the state; therefore, the set of all equilibria is the same across states. However, real people often have nonselfish mo-

[1] Several previous works such as the principal-agent problem with hidden information assumed public monitoring technology. In their environment, the central planner is able to detect whether an agent announces truthfully by observing ex-post public signals. See Salanié (2005) for a survey on the principal-agent problem with hidden information.

[2] Previous works in the literature on auction and implementation theory assumed that the central planner can utilize resource allocation to extract correct information by allowing agents to select an option from a menu of allocations and payoffs. See Krishna (2009) for a survey on auction theory and Maskin and Sjöström (2002) and Palfrey (2002) for surveys on implementation theory.

tives. Hence, the statement derived from this assumption could be useful only if it is robust against contamination of nonselfish motives.

This study considers the possibility in epistemology that some agents are honest, that is, they are motivated by an *intrinsic preference for honesty* as well as monetary interest. Many empirical and experimental studies indicate that human beings are not purely motivated by monetary payoffs but have intrinsic preferences for honesty. Abeler, Nosenzo, and Raymond (2019) provide a detailed meta-analysis (in which they use combined data from 90 studies involving more than 44,000 subjects across 47 countries) to show that subjects gave up a large fraction of potential gain from lying.[3]

However, our study allows the case in which an honest agent exists only exceptionally. We allow honest agents to be motivated mostly by monetary interest; the influences of preferences for honesty on decision making can be arbitrarily small. Furthermore, this study does not assume that agents expect the possibility that there exists an honest agent: we allow agents to have mutual knowledge that all agents are selfish (i.e., all agents know that all agents are selfish).

Despite these weaknesses in honesty, this study shows a very permissive result: the central planner can overcome the multiplicity of equilibria and elicit correct information from agents through unique Bayes Nash equilibrium (BNE) behavior if and only if "all agents are selfish" is not common knowledge.

In this study, the design of the payment rule has the following characteristics. First, each agent is required to announce not a state but a *distribution* of the state, while she (or he) is fully informed of the state; she can continuously fine-tune her announcement and payoff. Second, a selfish agent prefers to match her message with the other agents' messages. Third, an honest agent is driven to be *more honest* than a selfish agent. Due to these three characteristics, all agents come to expect the possibility that some of the other agents are driven to be more honest, which drives them into a tail-chasing competition toward honest reporting.

Specifically, we design the payment rule as the *quadratic scoring rule* (Brier, 1950), which aligns agents' payoffs with the distance between their messages. Hence, an agent's monetary payoff is maximized when she reports the average of the other agents' messages. The quadratic scoring rule is one of the standard mechanism design methods in partial implementation with asymmetric information.[4] This study suggests that this method is a powerful solution not only for partial

[3] Various works in behavioral economics and decision theory modeled preferences for honesty, such as a cost of lying (e.g., Ellingsen and Johannesson (2004); Kartik (2009)), a reputational cost (e.g., Mazar, Amir, and Ariely (2008)), and guilt aversion (e.g., Charness and Dufwenberg (2006)).

[4] See Cooke (1991) for a survey of scoring rules. For the applications to mechanism design, see for example Johnson, Pratt, and Zeckhauser (1990); Matsushima (1990, 1991, 1993, 2007); Aoyagi (1998); Miller, Pratt, Zeckhauser, and Johnson (2007). A number of studies extended the scoring rule of Brier (1950) to a setting in which a central planner collects information from a group of agents (e.g., Dasgupta and Ghosh (2013); Prelec (2004); Miller, Resnick, and Zeckhauser (2005); Kong and Schoenebeck (2019)). Previous

implementation but also for *unique implementation*.

The quadratic scoring rule may put each agent aside various nonselfish motives and prioritize her monetary interest to announce the same messages as those of other agents. However, as Abeler et al. (2019) pointed out, the intrinsic preference for honesty remains unexcluded in this case; therefore, honest agents still have an incentive to announce a little more honestly than selfish agents.

The equilibrium analysis of the game in the presence of behavioral agents and incomplete information itself has a long history. For example, Kreps, Milgrom, Roberts, and Wilson (1982) studied how the existence of behavioral agents changes the equilibria of finitely repeated games. Postlewaite and Vives (1987); Carlsson and Van Damme (1993); Morris and Shin (1998) studied how incomplete information shrinks the set of equilibria. These studies focused on the analysis of given games. In contrast, our focus is on the design of mechanisms, which takes full advantage of the potential existence of behavioral agents in higher-order beliefs.

Several studies have investigated the unique (or full) implementation of social choice functions, assuming existence of honest agents (Matsushima, 2008a,b; Dutta and Sen, 2012; Kartik, Tercieux, and Holden, 2014). In contrast to these works, this study does not make such an assumption—the only assumption we need is that "all agents are selfish" is not common knowledge.[5]

Just like the email game of Rubinstein (1989), this study contrasts the outcome under common knowledge and "almost common knowledge" of all agents' selfishness (in the sense that honest agents exist only in higher-order beliefs). As we assume that all agents are rational and can correctly distinguish these two, the possibility of honesty in epistemology drives all agents to tell the truth. The email game of Rubinstein (1989) demonstrates that "almost common knowledge" could lead us an unintuitive outcome. While our paper also demonstrates the vulnerability of the common knowledge assumption, the implication of this study is contrasting to Rubinstein's (1989). In an information elicitation mechanism, the intuitive outcome is truthtelling, and people can naturally expect that a truthful strategy profile is a focal point, while there are many equilibria under common knowledge of selfishness. We further show that, by carefully designing a "game" (mechanism), we can eliminate all the unwanted and unintuitive equilibria, whenever "all agents are selfish" is not common knowledge (while it could be "almost common knowledge"). Our result indicates that, when agents believe that others behave honestly "by default" and the mechanism nicely incentivizes agents to match their announcements, then it is difficult for agents to coordinate at a dishonest reporting.

This study is organized as follows. Section 2 presents the model, and Section 3 presents the main theorem (Theorem 1). Section 4 provides an example that outlines the logic behind Theorem 1

---

studies commonly assumed that all agents are selfish and, thus, suffered from the multiplicity of equilibria in a "single-question" setting in which the state is realized only once (as in our model).

[5] Matsushima (2020) showed, as an extension of this study, that all social choice functions are uniquely implementable in BNE if "all agents are selfish" is not common knowledge.

and Section 5 provides the proof of Theorem 1. Section 6 discusses issues such as application, mixed strategies, budget balancing, number of participants, and robustness against other behavioral motives. Section 7 extends Theorem 1 to the case in which each agent has ex-ante uncertainty in information access about which state occurs and is unknown to who else are informed agents. Section 8 considers the case in which agents have asymmetric information about the state. Section 9 concludes.

2 Model

This study investigates a situation in which a central planner attempts to correctly elicit information from multiple agents. Let $N = \{1,2,...,n\}$ denote the finite set of all agents, where $n \geq 2$. Let Ω denote the nonempty and finite sets of possible states. We assume *complete information about the state* across agents (in Sections 7 and 8, we discuss the case of incomplete information about the state). Each agent is informed of the true state $\omega \in \Omega$, whereas the central planner does not know it. Hence, the central planner attempts to design a mechanism that incentivizes these agents to make a truthful announcement.

From an epistemological perspective, we assume that agents are not always *selfish*; that is, they are not always concerned only about their monetary interests. Instead, agents could be *honest*, that is, motivated not only by monetary interest but also by an intrinsic preference for honesty. We assume *incomplete information concerning honesty* in that each agent knows whether they are selfish or honest, but the other agents are not informed of it. To formulate this incomplete information, we define the type space as follows, which is based on Bergemann and Morris (2005, 2013):

$$\Gamma \equiv (T_i, \pi_i, \theta_i)_{i \in N},$$

where $t_i \in T_i$ is agent i's type, $\theta_i : T_i \rightarrow \{0,1\}$ represents agent i's honesty, and $\pi_i : T_i \rightarrow \Delta(T_{-i})$ denotes agent i's belief about the other agents' types.[6] We assume that there exists a common prior $\pi \in \Delta(T)$ such that for all $t_i \in T_i$ and $t_{-i} \in T_{-i}$ such that $\sum_{t'_{-i} \in T_{-i}} \pi(t_i, t'_{-i}) > 0$, we have

$$\pi_i(t_{-i} \mid t_i) = \frac{\pi(t_i, t_{-i})}{\sum_{t'_{-i} \in T_{-i}} \pi(t_i, t'_{-i})}.$$

Each agent i knows her type t_i and the state ω, but does not know the other agents' types t_{-i}. Each agent is either selfish or honest: agent i is selfish (honest) if $\theta_i(t_i) = 0$ ($\theta_i(t_i) = 1$, respectively). More details will be subsequently explained.

[6]We denote by $\Delta(Z)$ the space of probability measures on the Borel field of a measurable space Z. We denote $Z \equiv \times_{i \in N} Z_i$, $Z_{-i} \equiv \times_{j \neq i} Z_j$, $z = (z_i)_{i \in N} \in Z$, and $z_{-i} = (z_j)_{j \neq i} \in Z_{-i}$.

The central planner designs a mechanism $G \equiv (M, x)$, where $M = \times_{i \in N} M_i$ denotes a message space, $x = (x_i)_{i \in N}$ denotes a payment rule, and $x_i : M \to R$ denotes the payment rule for agent i. Each agent i simultaneously announces a message $m_i \in M_i$ and obtains a monetary payment $x_i(m) \in R$ from the central planner, where we denote $m = (m_j)_{j \in N} \in M$.

We consider a class of indirect mechanisms in which each agent announces a probability distribution over states as the message, i.e., $M_i = \Delta(\Omega)$ for all $i \in N$. We write $m_i = \omega$ if $m_i(\omega) = 1$. A *strategy* of agent i is defined as $s_i : \Omega \times T_i \to M_i$, according to which agent i with type t_i announces the probability distribution over states $m_i = s_i(\omega, t_i) \in M_i = \Delta(\Omega)$ when the state $\omega \in \Omega$ occurs.

If agent i is selfish, her payoff is equal to her monetary payoff:

$$U_i(m; \omega, t_i, G) = x_i(m) \text{ if } \theta_i(t_i) = 0.$$

In contrast, if agent i is honest, she is motivated not only by monetary interest but also by an intrinsic preference for honesty.

$$U_i(m; \omega, t_i, G) = x_i(m) - c_i(m, \omega, t_i, G) \text{ if } \theta_i(t_i) = 1,$$

where $c_i(m, \omega, t_i, G) \in R$ denotes agent i's psychological cost. We assume that c_i represents the intrinsic preference for honesty. Specifically, for every $i \in N$, $\omega \in \Omega$, $m \in M$, and $\tilde{m}_i \in M_i$, if $\theta_i(t_i) = 1$ and $m_i(\omega) > \tilde{m}_i(\omega)$, then $c_i(m, \omega, t_i, G) \le c_i(\tilde{m}_i, m_{-i}, \omega, t_i, G)$, and

$$[x_i(\tilde{m}_i, m_{-i}) > x_i(m)] \Rightarrow [c_i(m, \omega, t_i, G) < c_i(\tilde{m}_i, m_{-i}, \omega, t_i, G)]. \tag{1}$$

Condition (1) implies that any honest agent feels guilty if she gains monetary payoffs from telling a lie. Hence, any honest agent strictly prefers making an announcement more honestly than the selfish types. In this study, we allow each agent's psychological cost to be arbitrarily small, even if this agent is honest: we make no assumption on the magnitude of psychological costs.

An example of psychological cost is given by

$$c_i(m, \omega, t_i, G) = \lambda_i \{1 - m_i(\omega)\},$$

where $\lambda_i > 0$. This example describes the preference for honesty with which an agent can save psychological cost by making an announcement more honestly.

Another example is given by

$$c_i(m, \omega, t_i, G) = \max[0, x_i(m) - x_i(\omega, m_{-i})] \lambda_i \{1 - m_i(\omega)\}$$

where $\lambda_i > 0$. In this example, the psychological cost depends crucially on the shape of the payment rule: how the lie changed the agent's monetary payoff influences the size of the psychological cost.

In both examples, by setting λ_i close to zero, we can consider the case in which the direct impact of the preference for honesty on an agent's decision-making can be arbitrarily small. In such a case, even honest agents are mostly motivated by monetary interests.

As implied by the latter example, we can also consider the case in which the direct impact of preference for honesty on an agent's decision-making is arbitrarily small compared with the impact of the lie on their monetary payoff. Importantly, as the latter example suggests, our model allows the case in which an honest agent incurs no psychological cost from telling a white lie (which has no influence on her monetary payoff). This study investigates Bayes Nash equilibrium in a game associated with a mechanism G. A strategy profile s is said to be a *Bayes Nash equilibrium* (BNE) if for every $\omega \in \Omega$, $i \in N$, $t_i \in T_i$, and $m_i \in M_i$,

$$E\left[U_i(s(\omega,t);\omega,t_i,G)\middle|\,\omega,t_i\right] \geq E\left[U_i(m_i,s_{-i}(\omega,t_{-i});\omega,t_i,G)\middle|\,\omega,t_i\right].$$

3 Main Theorem

We specify the payment rule $x = x^*$ as the following quadratic scoring rule: for every $i \in N$ and $m \in M$,

$$x_i^*(m) = -\sum_{j\neq i}\left[\sum_{\omega\in\Omega}\{m_i(\omega)-m_j(\omega)\}^2\right]$$

which describes the distance of agent i's message from the other agents' messages. From simple calculations, if s is a BNE in the game associated with x^*, then for every $i \in N$ and $(\omega,t_i) \in \Omega \times T_i$,

$$[\theta_i(t_i)=0] \Rightarrow \left[s_i(\omega,t_i) = E\left[\frac{\sum_{j\neq i} s_j(\omega,t_j)}{n-1}\middle|\,\omega,t_i\right]\right], \tag{2}$$

whereas

$$[\theta_i(t_i)=1] \Rightarrow \left[s_i(\omega,t_i)(\omega) = 1 \text{ or } s_i(\omega,t_i)(\omega) > E\left[\frac{\sum_{j\neq i} s_j(\omega,t_j)(\omega)}{n-1}\middle|\,\omega,t_i\right]\right]. \tag{3}$$

That is, any selfish agent mimics the average of the other agents' announcements in expectation, whereas any honest agent makes announcements more honestly than a selfish agent.

We define the *truthful strategy profile* s^* by

$$s_i^*(\omega,t_i) = \omega \text{ for all } i \in N \text{ and } (\omega,t_i) \in \Omega \times T_i,$$

according to which each agent i announces truthfully about the state irrespective of the state and type. We consider a necessary and sufficient condition under which the truthful strategy profile s^* is the unique BNE in the game associated with x^*; that is, the central planner succeeds in eliciting correct information about the state from the agents as unique equilibrium behavior.

We call a subset of type profiles $E \subset T \equiv \times_{i\in N} T_i$ an event. For convenience, for each event $E \subset T$, we write

$$\pi_i(E \mid t_i) = \pi_i(E_{-i}(t_i) \mid t_i),$$

where we denote $E_{-i}(t_i) \equiv \{t_{-i} \in T_{-i} | (t_i, t_{-i}) \in E\}$. Consider an arbitrary event $E \subset T$. Let

$$V_i^1(E) \equiv \{t_i \in T_i \mid \pi_i(E \mid t_i) = 1\}.$$

If $t_i \in V_i^1(E)$, then agent i knows the event E occurs. For each $k \geq 2$, let

$$V_i^k(E) \equiv \left\{ t_i \in T_i \middle| \pi_i \left(\mathop{\times}_{j\in N} V_j^{k-1}(E) \middle| t_i \right) = 1 \right\}.$$

If $t_i \in V_i^k(E)$, then agent i knows the event $\times_{j\in N} V_j^{k-1}(E)$ occurs. Note that, $V_i^k(E) \subset V_i^{k-1}(E)$ holds for all $k \geq 2$ because $t_i \notin V_i^{k-1}(E)$ implies $\pi_i(\times_{j\in N} V_j^{k-1}(E) \mid t_i) = \pi_i(\emptyset \mid t_i) = 0$, and therefore, $t_i \notin V_i^k(E)$. We define

$$V_i^\infty(E) \equiv \bigcap_{k=1}^{\infty} V_i^k(E).$$

An event $E \subset T$ is said to be *common knowledge* at $t \in T$ if

$$t \in \mathop{\times}_{i\in N} V_i^\infty(E).$$

Note that if E is common knowledge at $t \in T$, then

$$\pi_i \left(\mathop{\times}_{j\in N} V_j^\infty(E) \middle| t_i \right) = 1 \text{ for all } i \in N.$$

We denote by $E_i^* \subset T_i$ the set of agent i's types at which agent i is selfish, i.e.,

$$E_i^* = \{t_i \in T_i \mid \theta_i(t_i) = 0\}.$$

We denote the event that all agents are selfish by $E^* \equiv \times_{i\in N} E_i^*$.

Theorem 1. *The truthful strategy profile s^* is the unique BNE in the game associated with the* *quadratic scoring payment rule x^* if and only if*

$$\bigtimes_{i \in N} V_i^{\infty}(E^*) = \emptyset.$$

Section 4 introduces an example for providing intuition, and Section 5 shows a formal proof of Theorem 1.

Theorem 1 states that *all* agents (whether selfish or honest) will announce the state truthfully as unique BNE behavior if and only if "all agents are selfish" is not common knowledge. Hence, with the elimination of common knowledge of all agents' selfishness, the central planner can always succeed in eliciting correct information about the state from agents. We should regard this elimination as the minimal requirement of an epistemological potential that an agent cares about honesty. In fact, the success of correct elicitation holds even if "all agents are selfish" is mutual knowledge.

4 Example

The following characteristics of the quadratic scoring rule x^* are crucial for understanding Theorem 1. For simplicity of the arguments, we focus on the two-agent case.

(a) Each agent's message space is not the set of states but the set of probability distributions over states. Hence, an agent can continuously fine-tune their message and payment.

(b) Any selfish agent is incentivized to match her message with the other agent's message.

(c) Any honest agent is also incentivized to match her message with the other agent's message, but due to the intrinsic preference, she wants to behave slightly more honestly than the other agent.

(d) Suppose that agent 1 expects the possibility that agent 2 makes an announcement more honestly than what agent 2 expects about agent 1's announcement. Then, since agent 1 rationally expects that agent 2 attempts to announce more honestly, agent 1 with selfish type has an incentive to make the announcement more honestly than agent 2 initially expects. The same scenario holds even if agents 1 and 2 are replaced. This will be the driving force for a tail-chasing competition through which each agent announces more honestly than the other, reaching honest reporting by both.

Whenever agent i expects the possibility that the other agent $j \neq i$ is honest, then the supposition in (d) holds and agent i is driven to be more honest. However, the other agent j does not have to be honest: it is necessary and sufficient that agent i expects the possibility that the other agent j, whether selfish or honest, is driven to be more honest.

Consider the following example with a binary state space and a finite type space, where $n = 2$, $\Omega = \{0,1\}$, and $T_i = \{1,...,H\}$ for each $i \in \{1,2\}$. We assume that agent i is honest if and only if $t_i = 1$, that is, $E_i^* = \{2,...,H\}$. The message space of agent i is given by $M_i = [0,1]$, where $m_i \in [0,1]$ indicates the probability that state 1 ($\omega = 1$) occurs. The quadratic scoring rule is given by $x_1^*(m) = x_2^*(m) = -(m_1 - m_2)^2$.

We assume that the common prior is symmetric; that is, $\pi(h,h') = \pi(h',h)$ for all $(h,h') \in \{1,...,H\}^2$. Because the mechanism and agents are symmetric, we often refer to an agent with type h as a "type-h agent" without specifying their identity $i \in \{1,2\}$. We assume that the set of selfish types $E_i^* = \{2,\dots,H\}$ is path-connected in the sense that

$$\pi(h,h+1) > 0 \text{ for all } h \in \{2,\dots,H-1\}.$$

Without loss of generality, we assume that the true state is $\omega = 1$ (the analysis for the case of $\omega = 0$ is similar), and we drop it from the notation. The psychological cost for each agent i with honest type is given by $\lambda(1 - m_i)$, where $\lambda > 0$. Let $\bar{m}_j(t_i;s_j)$ be agent j's expected message conditional on agent i's type t_i:

$$\bar{m}_j(t_i;s_j) \equiv E\left[s_j(t_j) \mid t_i\right] = \sum_{h=1}^{H} \pi_i(h \mid t_i) s_j(h).$$

Then, agent i's best response against s_j is given by

$$BR_i(s_{-i},t_i) = \begin{cases} \bar{m}_j(t_i;s_j) & \text{if } t_i \in \{2,\dots,H\}; \\ \min\left\{\bar{m}_j(t_i;s_j) + \dfrac{\lambda}{2}, 1\right\} & \text{if } t_i = 1. \end{cases}$$

Hence, any honest agent is driven to be more honest than a selfish agent.

Case 1 First, consider the case in which the set of selfish types is disconnected from the honest type, that is,

$$\pi(1,h) = 0 \text{ for all } h \in \{2,\dots,H\}.$$

Any selfish agent expects that the other agent is selfish with certainty, and any honest agent expects that the other agent is honest with certainty.

When $t = (1,1)$ is realized, the best response of each (honest) agent $i \in \{1,2\}$ is given by $s_i(1) = \min\{s_j(1) + \lambda/2, 1\}$; that is, the preference for honesty drives each agent to select a message that is slightly more honest than the other. Clearly, whenever s is a BNE, $s_1(1) = s_2(1) = 1$ must be satisfied.

In contrast, an equilibrium strategy can take any value when $h \in \{2,...,H\}$. As long as there exists a constant $p \in [0,1]$ such that

$$s_i(h) = p \text{ for all } i \in N \text{ and } h \in \{2,\ldots,H\},$$

it is a BNE. Hence, there are infinitely many BNEs in which any selfish agent may tell a lie. Clearly, we fail to elicit the correct state as a unique BNE in Case 1.

Case 2 Consider the case in which, unlike Case 1, the set of selfish types is connected with the honest type in a minimal sense such that there exists $h \in \{2,...,H\}$ with $\pi(1,h) > 0$. For simplicity, we assume that $h = 2$, that is,

$$\pi(1,2) > 0.$$

It is easy to see that the same argument holds even if we replace type 2 with any $h \in \{3,...,H\}$.

Because of higher-order reasoning, this minimal connection drastically changes the set of BNEs as follows. Clearly, a type-1 (honest) agent is driven to be more honest. The minimal connection implies that a type-2 agent expects that the other agent may be type-1 with a positive probability. Since a type-2 agent would like to match her message with the other agent (who could be type-1), she is also driven to be more honest. Similarly, a type-3 agent expects that the other agent may be type-2 with a positive probability and, thus, is driven to be more honest. We can iterate this argument and verify that any agent, whether selfish or honest, is driven to be more honest, that is, attempts to send a more honest message than the other. This structure of best responses immediately leads us to the uniqueness of BNE, where all agents report truthfully.

Note that this uniqueness holds even if both agents' selfishness is mutual knowledge. As long as $t_1 \geq 3$ and $t_2 \geq 3$, each agent does not expect that the other agent may be honest. However, the aforementioned higher-order reasoning will guide any agent to send a more truthful message, which drastically shrinks the set of BNE. As long as there is no common knowledge of both agents' selfishness, this logic always functions and the uniqueness of the BNE is guaranteed.

Case 1 corresponds to situations in which all selfish types completely eliminate associations with honest types. In this case, unique information elicitation is impossible. By contrast, as in Case 2, if there is at least one selfish type who expects even a little (possibly indirect) influence of an honest type, then unique information elicitation is achievable. The driving force behind this phenomenon is not that more people become honest but that selfish people do not rule out the existence of honest agents from their epistemological considerations.

5 Proof of Theorem 1

It is clear from (2) and (3) that s^* is a BNE; thus, it suffices to show the uniqueness. Suppose that s is a BNE. Fix $\omega \in \Omega$ arbitrarily. Let

$$\alpha \equiv \min_{(i,t_i)} s_i(\omega, t_i)(\omega),$$

and

$$\tilde{T}_i \equiv \{t_i \in T_i \mid s_i(\omega, t_i)(\omega) = \alpha\}$$

for each $i \in N$. Suppose that

$$\times_{i \in N} V_i^\infty(E^*) = \emptyset.$$

From the definition of common knowledge, this supposition is equivalent to

$$V_i^\infty(E^*) = \emptyset \text{ for all } i \in N.$$

Toward a contradiction, suppose that $\alpha < 1$, i.e., there exists an agent $i \in N$ and type $t_i \in T_i$ that does not adopt the truthful strategy. Note from (2) and (3) that any honest agent prefers making announcements more honestly than selfish agents, implying that no honest type belongs to $\tilde{T}_i$; i.e., $\tilde{T}_i \subset E_i^*$.

Consider an arbitrary $i \in N$ and $t_i \in \tilde{T}_i$. From (2) and (3), α equals the average of the other agents' announcements on ω in expectation but not greater than any announcement. Hence, type t_i expects that any other agent $j \neq i$ announces $m_i(\omega) = \alpha$, that is,

$$\pi_i\left(\times_{j \in N} \tilde{T}_j \middle| t_i\right) = 1.$$

This, along with (2) and (3), implies that agent i with type t_i expects that the other agents are surely selfish, that is,

$$\pi_i(E^* \mid t_i) = 1.$$

Hence, we have

$$\tilde{T}_i \subset V_i^1(E^*).$$

Moreover, because

$$\pi_i\left(\underset{j\in N}{\times} V_j^1(E^*)\middle| t_i\right) \geq \pi_i\left(\underset{j\in N}{\times} \tilde{T}_j\middle| t_i\right) = 1,$$

we have $\pi_i(\times_{j\in N}V_j^1(E^*) \mid t_i) = 1$, that is,

$$\tilde{T}_i \subset V_i^2(E^*).$$

Similarly, we have

$$\tilde{T}_i \subset V_i^k(E^*) \text{ for all } k \geq 2.$$

Hence, we have

$$\tilde{T}_i \subset V_i^\infty(E^*),$$

which however contradicts the supposition that $V_i^\infty(E^*) = \emptyset$. Hence, we conclude that $\alpha = 1$ or, equivalently, $s_i(\omega, t_i) = \omega$ for all $\omega \in \Omega$. Accordingly, $s = s^*$ must be the case in any BNE. Here, we have proved the "if" part of Theorem 1.

Fix an arbitrary $\omega' \neq \omega$. We specify a strategy profile s^+ as follows: for every $i \in N$ and $t_i \in T_i$,

$$s_i^+(\omega, t_i) = \omega \quad \text{if } t_i \notin V_i^\infty(E^*);$$
$$s_i^+(\omega, t_i) = \omega' \quad \text{if } t_i \in V_i^\infty(E^*),$$

and

$$s_i^+(\tilde{\omega}, t_i) = s_i^*(\tilde{\omega}, t_i) \quad \text{for all } \tilde{\omega} \neq \omega.$$

It is clear from (2) and the previous argument that s^+ is a BNE, and $s^+ \neq s^*$ whenever $V_i^\infty(E^*) \neq \emptyset$ for some $i \in N$. Hence, we have proven the "only-if" part of Theorem 1. □

6 Discussion

6.1 Application: Blockchain and Oracle Problem

This study assumed that the central planner has the power to force payments according to the predetermined mechanism (quadratic scoring rule). However, a companion work (Matsushima and Noda, 2020) points out that the argument in this study does not depend on the presence of such a central planner or the court; without external coercion, we can automate and self-enforce the monetary payment rule within the scope of current digital technology. That is, by using digital currencies, the message-contingent monetary payment rule can be computer-programmed as a so-called *smart contract* and deployed on a blockchain such as Ethereum. However, in this case, we face the problem of how to incentivize agents to input correct information into the smart contract. This issue is called the *oracle problem* in the blockchain literature. This problem is regarded as one of the most serious problems that hinders the effective use of smart contracts. Matsushima and Noda (2020) show that this study's theorem provides a new and promising direction to solve this problem.

6.2 Mixed Strategies

Thus far, we have considered only pure strategy BNE. However, we can directly use the same logic for the uniqueness of the mixed strategy BNE. Because of the quadratic scoring rule, irrespective of whether the other agents' strategies are mixed or pure, any selfish agent prefers announcing the same distribution as the other agents' announcements in expectation, whereas any honest agent prefers announcing more honestly than a selfish agent. The resultant tail-chasing competition eliminates any unwanted BNE, including mixed equilibria.

6.3 Budget Balance

The quadratic scoring rule x^* does not balance the budget. In fact, in the two-agent case, it is difficult to find an alternative rule that induces unique information elicitation x^* and balances the budget. In contrast, in the three-or-more-agent case, it is easy to check that the following payment rule x^+ induces unique information elicitation and satisfies the budget-balance property: for every $i \in N$ and $m \in M$,

$$x_i^+(m) = x_i^*(m) + r_i(m_{-i}),$$

where

$$r_i(m_{-i}) \equiv \frac{1}{n-2} \sum_{i' \neq i, j \neq i, i' \neq j} \left[\sum_{\omega \in \Omega} \{m_{i'}(\omega) - m_j(\omega)\}^2 \right].$$

6.4 Number of Participants

Theorem 1 holds irrespective of the number of agents participating in the central planner's problem. However, informally, the restrictiveness of the necessary and sufficient conditions depends on this number. In other words, the more agents participate in the problem, the less likely it is that "all agents are selfish" is common knowledge. If the number of participants is limited, the central planner should recruit informed people from a wider range. If an agent is selfish and believes that she and other agents are alike, the agent is likely to expect common knowledge of all agents' selfishness. In such a case, the set of selfish types becomes disconnected from the honest type and unique information elicitation may fail. If participants have diverse backgrounds, then selfish agents may expect that the others have different preferences and beliefs, and therefore, truthful messages could be induced.

6.5 Other Behavioral Motives

This study assumed that there exist only two categories of agents: selfish agents and honest agents. However, there could be more diverse irrational motives, such as "always tell a lie" and "always announce a fixed message." If we explicitly consider these motives, we can no longer show that a selfish agent is attracted to announce a literally truthful message. Nevertheless, the equilibrium messages are attracted to somewhere close to truth-telling whenever these motives are not as important as honesty, and the central planner can identify the true state by checking whether agents' messages are attracted by a certain message.

Abeler et al. (2019) empirically and experimentally studied intrinsic preferences for honesty. They show that preferences for honesty are the main motivation in a wide range of observed behaviors. Their result supports the assumption that each agent is either selfish or honest. However, the dominant motive may depend on the context, and the central planner had better select agents from a population that has little to do with her purpose.

We also point out that agents' behavioral motivations could be influenced by the scale of the payment rules. More specifically, when a central planner sets up an excessively small-scale payment rule, agents may disrespect the mechanism and may not be motivated to behave honestly. In such a case, the intrinsic preference for honesty could be weakened, and other behavioral motivates could significantly affect agents' behavior. Recall that Theorem 1 itself assumes no condition on the scale of the quadratic scoring rule. Accordingly, if the central planner can tune the scale to enhance the preference for honesty, then the payment scale could be selected in such a way that honesty dominates the other behavioral motives.

7 Uncertainty in Information Access

Thus far, we have assumed that the set of agents who are informed about the state and participate in the elicitation mechanism is common knowledge. However, in many real-world problems, each agent may not possess complete information about other participants. Since our aim is to investigate how the lack of common knowledge influences the (im-)possibility of implementation, we also need to explore how the uncertainty about the set of participants would affect the conclusion.

In this section, we consider a case in which the set of informed agents could be uncertain ex ante. Whether an agent i is informed or not is determined by her type t_i (and therefore, it is private information). When agent i is uninformed, she does not participate in the mechanism, and therefore, does not submit a message to the mechanism. Each agent does not know whether the other agents are informed or not, but form a belief about it based on her own type. Accordingly, an agent's (non-)participation may not be common knowledge among agents.

We modify the type space as follows:

$$\Gamma \equiv (T_i, \pi_i, \theta_i, \eta_i)_{i \in N}.$$

Here, we additionally introduced $\eta_i : T_i \to \{0,1\}$, which indicates whether agent i is informed or not. Agent i is *informed* (*uninformed*) if $\eta_i(t_i) = 1$ ($\eta_i(t_i) = 0$, respectively). Agent i with $\eta_i(t_i) = 0$ does not know which state occurs, and therefore, does not participate in the mechanism. Since we ignore incentives of uninformed agents, when we refer to an agent as either selfish or honest, it implicitly implies that she is informed.

When an agent believes that she is the only participant of the mechanism, clearly mutual monitoring does not work, and unique implementation is impossible. Hence, we assume that any informed agent expects the possibility that there exist other agents who are informed: for every $i \in N$ and $t_i \in T_i$, if $\eta_i(t_i) = 1$, we have

$$\pi_i\left(\left\{t_{-i} \in T_{-i} \mid \exists j \in N \setminus \{i\} \text{ s.t. } \eta_j(t_j) = 1\right\} \big| t_i\right) > 0.$$

We assume that uninformed agents do not submit a message to the mechanism. Accordingly, the input of the mechanism (or payment rule) is a profile of messages announced by informed agents. Accordingly, the payment rule x is defined as

$$x : \bigcup_{H \subset 2^N} \left(\times_{i \in H} M_i \right) \to R^N,$$

that is, the central planner pays $x_i(m_H) \in R$ to agent i if the set of informed agents is H, i.e., $H = H(t) \equiv \{j \in N \mid \eta_j(t_j) = 1\}$, and they announce a message profile $m_H \in \times_{j \in H} M_j$.

The definition of the utility function is the same as the definition of Section 2 of this study. Since we assume that uninformed agents do not participate in the mechanism, $s_i(\omega, t_i) = \emptyset$ must be the case if $\eta_i(t_i) = 0$. A strategy profile s is a *Bayes Nash equilibrium* (BNE) in the game associated with x if for every $\omega \in \Omega$, $i \in N$, $t_i \in T_i$, and $m_i \in M_i$, whenever $\eta_i(t_i) = 1$, we have

$$E\left[U_i(s(\omega,t);\omega,t_i,G) \middle| \omega,t_i\right] \geq E\left[U_i(m_i,s_{-i}(\omega,t_{-i});\omega,t_i,G) \middle| \omega,t_i\right].$$

We define the truthful strategy profile s^{**}: for every $i \in N$ and $(\omega,t_i) \in \Omega \times T_i$,

$$s_i^{**}(\omega,t_i)(\omega) = 1 \text{ if } \eta_i(t_i) = 1.$$

Our aim is to construct a mechanism that achieves s^{**} as a unique BNE.

Parallel to the quadratic payment rule we have studied in the previous sections, we define the quadratic payment rule for this environment, x^{**}, in the following manner: for every $i \in N$ and $m \in M$,

$$x_i^{**}(m_H) = -\sum_{j \in H} \left[\sum_{\omega \in \Omega} \{m_i(\omega) - m_j(\omega)\}^2 \right] \text{ if } i \in H$$

and

$$x_i^{**}(m_H) = 0 \text{ if } i \notin H.$$

If s is a BNE in the game associated with x^{**}, for every $i \in N$ and $(\omega,t_i) \in \Omega \times T_i$, we have

$$[\theta_i(t_i) = 0 \text{ and } \eta_i(t_i) = 1] \Rightarrow \left[s_i(\omega,t_i) = E\left[\frac{\sum_{j \in H(t)\setminus\{i\}} s_j(\omega,t_j)}{|H(t)|-1} \middle| \omega,t_i \right] \right], \tag{4}$$

and

$$\begin{aligned} &[\theta_i(t_i) = 1 \text{ and } \eta_i(t_i) = 1] \\ &\Rightarrow \left[s_i(\omega,t_i)(\omega) = 1 \text{ or } s_i(\omega,t_i) > E\left[\frac{\sum_{j \in H(t)\setminus\{i\}} s_j(\omega,t_j)}{|H(t)|-1} \middle| \omega,t_i \right] \right]. \end{aligned} \tag{5}$$

That is, any selfish agent mimics the average of the other informed agents' announcements in expectation, while any honest agent announces more honestly than selfish agents.

We show that s^{**} is the unique BNE if and only if "all participants are selfish" is not common knowledge across all participants. First, we formally define such an event. We define $T_i^{\dagger}$ as the set of agent i's types with which agent i is uninformed:

$$T_i^{\dagger} \equiv \{t_i \in T_i \mid \eta_i(t_i) = 0\}.$$

Consider an arbitrary event $E \subset T$. We define

$$\bar{V}_i^1(E) \equiv \{t_i \in T_i \setminus T_i^\dagger \mid \pi_i(E \mid t_i) = 1\}.$$

For each $k \geq 2$, we define

$$\bar{V}_i^k(E) \equiv \left\{ t_i \in T_i \setminus T_i^\dagger \,\middle|\, \pi_i \left(\times_{j \in N} (\bar{V}_j^{k-1}(E) \cup T_j^\dagger) \,\middle|\, t_i \right) = 1 \right\}.$$

In plain words, $\bar{V}_i^k(E)$ is the set of agent i's types with which agent i is informed and knows that for every agent j, either $\bar{V}_j^{k-1}(E)$ or $T_j^\dagger$ occurs. Since $t_i \notin \bar{V}_i^{k-1}(E)$ and $t_i \notin T_i^\dagger$ implies $\pi_i(\times_{j \in N}(\bar{V}_j^{k-1}(E) \cup T_j^\dagger) \mid t_i) = 0$, $\bar{V}_i^k(E) \subset \bar{V}_i^{k-1}(E)$ for all $k \geq 2$. We define

$$\bar{V}_i^\infty(E) \equiv \bigcap_{k=1}^{\infty} \bar{V}_i^k(E).$$

An event $E \subset T$ is said to be *common knowledge across informed agents* at $t \in T$ if

$$t \in \times_{i \in N} \bar{V}_i^\infty(E).$$

Note that if E is common knowledge across informed agents at t, then

$$\pi_i \left(\times_{j \in N} (\bar{V}_j^\infty(E) \cup T_j^\dagger) \,\middle|\, t_i \right) = 1 \text{ for all } i \in N.$$

Theorem 2. *The truthful strategy profile s^{**} is the unique BNE in the game associated with x^{**} if and only if the following condition is satisfied:*

$$\times_{i \in N} \bar{V}_i^\infty(E^*) = \emptyset.$$

Proof. See the Appendix. □

Theorem 2 implies that the central planner can elicit correct information from all informed agents as unique BNE behavior if and only if "any participant is selfish" not common knowledge across all participants. Theorem 2 suggests that ex-ante uncertainty about who else are informed plays a great impact on incentivizing an informed agent to announce truthfully. For instance, if $H(t)$ is a singleton and the informed agent knows this fact ex ante, this agent is never motivated to tell the truth whenever she is selfish. On the other hand, with the condition in Theorem 2, if she does not know this fact ex ante, she is willing to tell the truth as unique equilibrium behavior, even when she is selfish.

8 Asymmetric Information

Thus far, we have assumed symmetric information in that agents access the same information channel concerning the state. The proofs of Theorems 1 and 2 rely on this assumption: the quadratic payment rule incentivizes each selfish agent to match the message with the other agents and each honest type to announce more honestly than selfish types, which creates the tail-chasing competition towards all agents' honest reporting. However, we cannot directly apply this proof procedure to asymmetric information environments where each agent only observes partial information about the state as her private signal because of the lack of mutual monitoring.

This section considers an extension of the quadratic scoring payment rule to show the possibility of unique information elicitation in an asymmetric information environment. In this section, we assume that each agent $i \in N$ only observes a private signal $\omega_i \in \Omega_i$. For simplicity of arguments, we assume that the space of private signals is binary, i.e., $\Omega_i = \{0, 1\}$. A state is defined as a profile of agents' private signals, $\Omega = \times_{i \in N} \Omega_i$, where we denote its generic element as $\omega = (\omega_i)_{i \in N} \in \Omega$. We denote by $p_{i,j}(\cdot \mid \omega_i) : \Omega_j \to [0, 1]$ the probability distribution over Ω_j conditional on ω_i. We assume that agents' private signals are positively correlated, and for any two agents $i, j \in N$, we have $p_{i,j}(1 \mid 1) > 1/2$ and $p_{i,j}(0 \mid 0) > 1/2$.

Differently from the previous sections, we focus on a class of mechanisms where the central planner requires each agent to announce a bundle of multiple sub-messages at once. Fix an arbitrary positive integer $H > 0$. Let $M_i = \times_{h=0}^{H} M_i^h$ and

$$M_i^h = \Delta(\Omega_i) \text{ for all } h \in \{0, 1, ..., H\}$$

where we denote $m_i = (m_i^h)_{h=0}^{H}$ and $m_i^h \in M_i^h$ for each $h \in \{0, 1, ..., H\}$. Each agent i reports $H + 1$ sub-messages at once, which reveals information about her private signal ω_i. At each hth sub-message, agent i announces a distribution on Ω_i, $m_i^h \in \Delta(\Omega_i)$. Since the space of private signals is binary, we can denote $m_i^h = m_i^h(1) \in [0, 1]$.

A strategy of agent i is defined as $s_i : \Omega_i \times T_i \to M_i$, according to which agent i with private signal ω_i and type t_i announces $s_i(\omega_i, t_i) \in M_i$. Let $s_i = (s_i^h)_{k=0}^{H}$, $s_i^h : \Omega_i \times T_i \to M_i^h$, and $s_i(\omega_i, t_i) = (s_i^h(\omega_i, t_i))_{h=0}^{H}$, where $s_i^h(\omega_i, t_i) \in M_i^h$ denotes agent i's hth sub-message.

The definition of the type space, $\Gamma = (T_i, \pi_i, \theta_i)_{i \in N}$, and events $V_i^k(E)$ and E^* are unchanged, but we assume that T is finite. The structure of agents' utility functions is also similar, but we need to update the specification of the psychological cost functions because we modified message spaces. Each agent i's payoff function $U_i(\,\cdot\,; \omega_i, t_i, G) : M \to R$ is defined as

$$U_i(m; \omega_i, t_i, G) = x_i(m) \text{ if } \theta_i(t_i) = 0,$$

and

$$U_i(m;\omega_i,t_i,G) = x_i(m) - c_i(m,\omega_i,t_i,G) \text{ if } \theta_i(t_i) = 1.$$

Here, we assume that the psychological cost function $c_i(\cdot,\omega_i,t_i,G) : M \to R$ satisfies the following condition: for every $\omega_i \in \Omega_i$, $m \in M$, $\tilde{m}_i \in M_i$, and $h \in \{0,...,H\}$,

$$\left[\theta_i(t_i) = 1, m_i^{-h} = \tilde{m}_i^{-h}, m_i^h(\omega_i) > \tilde{m}_i^h(\omega_i), \text{ and } x_i(\tilde{m}_i,m_{-i}) > x_i(m)\right]$$
$$\Rightarrow \left[c_i(\tilde{m}_i,m_{-i},\omega_i,t_i,G) > c_i(m,\omega_i,t_i,G)\right].$$

Similar to the model described in Section 2, we assume that honest agents incur larger psychological costs if they tell less truthful messages. We further assume that $c_i(m,\omega_i,t_i,G)$ is convex in m_i.

A strategy profile s is said to be a *Bayes Nash equilibrium* (BNE) in the game associated with the mechanism G if, for every $i \in N$, $\omega_i \in \Omega_i$, $t_i \in T_i$, and $m_i \in M_i$,

$$E[U_i(s_i(\omega_i,t_i),m_{-i};\omega_i,t_i,G) \mid \omega_i,t_i,s_{-i}] \geq E[U_i(m_i,m_{-i};\omega_i,t_i,G) \mid \omega_i,t_i,s_{-i}].$$

For each distinct $i, j \in N$, we construct a part of the payment rule $\hat{x}_i^j : M_i \times M_j \to R$ as a combination of multiple quadratic scoring rules:

$$\hat{x}_i^j(m_i,m_j) \equiv -\left(m_i^0 - \frac{1}{2}\right)^2 - \sum_{h=1}^{H}\left\{m_i^h - I_j(m_j^{h-1})\right\}^2.$$

where the function $I_i : [0,1] \to \{0,1/2,1\}$ is specified as follows:

$$I_i(p) \equiv \begin{cases} 0 & \text{if } p < 1/2, \\ 1/2 & \text{if } p = 1/2, \\ 1 & \text{if } p > 1/2. \end{cases}$$

Agent i's payment rule $\hat{x}_i : M \to R$ is defined as the sum of $(\hat{x}_i^j)_{j\neq i}$:

$$\hat{x}_i(m) \equiv \sum_{j\in N\setminus\{i\}} \hat{x}_i^j(m_i,m_j).$$

Note that, the payment rule $\hat{x}_i$ is strictly concave in m_i. Since the psychological cost function c_i is assumed to be convex in m_i^h, both selfish and honest agents have a unique best response.

Based on the function I_i, the central planner interprets agent i's hth sub-message $m_i^h \in [0,1]$ in the following manner:

(i) If $I_i(m_i^h) = 0$, then the central planner regards $\omega_i = 0$.

(ii) If $I_i(m_i^h) = 1$, then the central planner regards $\omega_i = 1$.

(iii) If $I_i(m_i^h) = 1/2$, then the central planner defers a decision.

Accordingly, if agent i sends a message $m_i^h \neq 1/2$, then the central planner can infer agent i's private signal ω_i. We say that m_i^h is *informative* if $I_i(m_i^h) = \omega_i$.

The following theorem shows that even in the asymmetric information environment, by observing the Hth messages, the central planner can successfully elicit correct private information from all agents through their unique BNE behaviors, whenever "all agents are selfish" not common knowledge.

Theorem 3. *There exists the unique BNE s in the game associated with the payment rule $\hat{x}$. Furthermore, if "all agents are selfish" is not common knowledge, i.e.,*

$$\times_{i \in N} V_i^{\infty}(E^*) = \emptyset,$$

then there exists a positive integer K such that whenever $H \geq K$, the Hth message profile in the unique BNE s correctly informs the state, i.e.,

$$I_i(s_i^H(\omega_i, t_i)) = \omega_i \textit{ for all } i \in N,\ \omega_i \in \Omega_i, \textit{ and } t_i \in T_i.$$

Proof. We define

$$\begin{aligned} T_i^h(0, \omega_i, s_i) &\equiv \{t_i \in T_i \mid I_i(s_i^h(\omega_i, t_i)) = 0\}, \\ T_i^h(1/2, \omega_i, s_i) &\equiv \{t_i \in T_i \mid I_i(s_i^h(\omega_i, t_i)) = 1/2\}, \text{ and} \\ T_i^h(1, \omega_i, s_i) &\equiv \{t_i \in T_i \mid I_i(s_i^h(\omega_i, t_i)) = 1\}. \end{aligned}$$

According to an iterative elimination method from the zeroth sub-messages to the Hth sub-messages, the BNE strategy profile s is uniquely determined according to the following steps.

Step 0 Because of the specification of $\hat{x}$, if agent i is selfish, she chooses m_i^0 to maximize $-(m_i^0 - 1/2)^2$, and therefore, $s_i^0(\omega_i, t_i) = 1/2$. Accordingly, $I_i(s_i^0(\omega_i, t_i)) = 1/2$ if $t_i \in E_i^*$. If she is honest, she maximizes $-(m_i^0 - 1/2)^2$ minus her psychological cost, which uniquely determines $s_i^0(\omega_i, t_i)$, where $s_i^0(0, t_i) < 1/2$ and $s_i^0(1, t_i) > 1/2$. Accordingly, we have $I_i(s_i^0(\omega_i, t_i)) = \omega_i$ if $t_i \notin E_i^*$. Since all agents are either selfish or honest, no agent reports an untruthful message, i.e., $I(s_i^0(\omega_i, t_i)) \notin \{1/2, \omega_i\}$. Hence, we have the following:

$$\begin{aligned} T_i^0(1/2, \omega_i, s_i) &= E_i^*, \text{ and} \\ T_i^0(\omega_i, \omega_i, s_i) &= T_i \setminus T_i^0(1/2, \omega_i, s_i). \end{aligned}$$

Step $h \geq 1$ Suppose that for every $i \in N$, $\omega_i \in \Omega_i$, and $h' \in \{0, ..., h-1\}$, we have

$$T_i^{h'}(1/2, \omega_i, s_i) = V_i^{h'}(E^*), \text{ and}$$
$$T_i^{h'}(\omega_i, \omega_i, s_i) = T_i \setminus T_i^{h'}(1/2, \omega_i, s_i),$$

where we denote $V_i^0(E^*) = E_i^*$. If agent i is selfish and expects that for every agent $j \neq i$, we have $t_j \in T_j^{h-1}(\omega_j, \omega_j, s_j)$ with certainty, she believes that $I_j(s_j^{h-1}(\omega_j, t_j)) = 1/2$. Hence, she maximizes $-(m_i^h - 1/2)^2$, and therefore, reports $s_i^h(\omega_i, t_i) = 1/2$. If agent i is selfish and expects that there exists $j \neq i$ such that $t_j \notin T_j^{h-1}(\omega_j, \omega_j, s_j)$ with a positive probability, by the induction hypothesis, $t_j \in T_j^{h-1}(1/2, \omega_j, s_j)$ must be the case. Hence, agent i maximizes a convex combination of $-(m_i^h - 1/2)^2$ and $-(m_i^h - I_j(\omega_j))^2$, and the probability weight on the latter term is positive. Accordingly, $s_i^h(\omega_i, t_i)$ is uniquely determined as a maximizer of the convex combination, and it must be informative, i.e., $I_i(s_i^h(\omega_i, t_i)) = \omega_i$. If agent i is honest, agent i maximizes a convex combination of $-(m_i^h - 1/2)^2$ and $-(m_i^h - I_j(\omega_j))^2$ minus her psychological cost. Hence, $s_i^h(t_i, \omega_i)$ is uniquely determined as a maximizer of it, and it must be informative, i.e., $I_i(s_i^h(\omega_i, t_i)) = \omega_i$. Accordingly, we have the following:

$$T_i^h(1/2, \omega_i, s_i) = V_i^h(E^*), \text{ and}$$
$$T_i^h(\omega_i, \omega_i, s_i) = T_i \setminus T_i^h(1/2, \omega_i, s_i).$$

From the above-mentioned steps, by mathematical induction, we have

$$T_i^H(1/2, \omega_i, s_i) = V_i^H(E^*), \text{ and}$$
$$T_i^H(\omega_i, \omega_i, s_i) = T_i \setminus T_i^H(1/2, \omega_i, s_i).$$

Since T is finite, there exists K such that

$$V_i^\infty(E^*) \equiv \bigcap_{k=1}^{K} V_i^k(E^*).$$

Since $V_i^\infty(E^*) = \emptyset$, we have

$$V_i^k(E^*) = \emptyset \text{ for all } k \geq K.$$

Hence, if $H \geq K$, we have

$$T_i^H(1/2, \omega_i, s_i) = \emptyset \text{ and } T_i^H(\omega_i, \omega_i, s_i) = T_i,$$

that is,

$$I_i(s_i^H(\omega_i, t_i)) = \omega_i \text{ for all } i \in N,\ \omega_i \in \Omega_i,\ \text{and}\ \ t_i \in T_i.$$

□

9 Concluding Remarks

We investigated the unique information elicitation in which the central planner uses only payment rules and agents are either selfish or honest. We proved that despite a severe lack of incentive devices availability, the central planner could elicit correct information from informed agents through their unique BNE behavior if and only if "all agents are selfish" is not common knowledge.

It is an important future research to investigate the case of asymmetric information concerning the state with general environments. Do quadratic scoring rules function? If not, what is the alternative design that generally solves unique information elicitation? Is the exclusion of common knowledge of all agents' selfishness generally sufficient? These questions represent only the tip of the iceberg but could include new theoretical substances beyond the scope of this study.

References

ABELER, J., D. NOSENZO, AND C. RAYMOND (2019): "Preferences for Truth-Telling," *Econometrica*, 87, 1115–1153.

AOYAGI, M. (1998): "Correlated Types and Bayesian Incentive Compatible Mechanisms with Budget Balance," *Journal of Economic Theory*, 79, 142–151.

BERGEMANN, D. AND S. E. MORRIS (2005): "Robust Mechanism Design," *Econometrica*, 1771–1813.

——— (2013): *An Introduction to Robust Mechanism Design*, Foundations and Trends in Microeconomics, Now Publishers Inc.

BRIER, G. W. (1950): "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78, 1–3.

CARLSSON, H. AND E. VAN DAMME (1993): "Global Games and Equilibrium Selection," *Econometrica*, 989–1018.

CHARNESS, G. AND M. DUFWENBERG (2006): "Promises and Partnership," *Econometrica*, 74, 1579–1601.

COOKE, R. M. (1991): *Experts in Uncertainty: Opinion and Subjective Probability in Science*, Oxford University Press on Demand.

DASGUPTA, A. AND A. GHOSH (2013): "Crowdsourced Judgement Elicitation with Endogenous Proficiency," in *Proceedings of the 22nd International Conference on World Wide Web*, 319–330.

DUTTA, B. AND A. SEN (2012): "Nash Implementation with Partially Honest Individuals," *Games and Economic Behavior*, 74, 154–169.

ELLINGSEN, T. AND M. JOHANNESSON (2004): "Promises, Threats and Fairness," *The Economic Journal*, 114, 397–420.

JOHNSON, S., J. W. PRATT, AND R. J. ZECKHAUSER (1990): "Efficiency Despite Mutually Payoff-Relevant Private Information: The Finite Case," *Econometrica*, 873–900.

KARTIK, N. (2009): "Strategic Communication with Lying Costs," *The Review of Economic Studies*, 76, 1359–1395.

KARTIK, N., O. TERCIEUX, AND R. HOLDEN (2014): "Simple Mechanisms and Preferences for Honesty," *Games and Economic Behavior*, 83, 284–290.

KONG, Y. AND G. SCHOENEBECK (2019): "An Information Theoretic Framework for Designing Information Elicitation Mechanisms that Reward Truth-Telling," *ACM Transactions on Economics and Computation (TEAC)*, 7, 1–33.

KREPS, D. M., P. MILGROM, J. ROBERTS, AND R. WILSON (1982): "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma," *Journal of Economic theory*, 27, 245–252.

KRISHNA, V. (2009): *Auction Theory*, Academic Press.

MASKIN, E. AND T. SJÖSTRÖM (2002): "Implementation Theory," *Handbook of Social Choice and Welfare*, 1, 237–288.

MATSUSHIMA, H. (1990): "Dominant Strategy Mechanisms with Mutually Payoff-Relevant Private Information and with Public Information," *Economics Letters*, 34, 109–112.

——— (1991): "Incentive Compatible Mechanisms with Full Transferability," *Journal of Economic Theory*, 54, 198–203.

——— (1993): "Bayesian Monotonicity with Side Payments," *Journal of Economic Theory*, 59, 107–121.

——— (2007): "Mechanism Design with Side Payments: Individual Rationality and Iterative Dominance," *Journal of Economic Theory*, 133, 1–30.

——— (2008a): "Behavioral Aspects of Implementation Theory," *Economics Letters*, 100, 161–164.

——— (2008b): "Role of Honesty in Full Implementation," *Journal of Economic Theory*, 139, 353–359.

——— (2020): "Implementation, Honesty and Common Knowledge," Working Paper.

MATSUSHIMA, H. AND S. NODA (2020): "Mechanism Design with Blockchain Enforcement," Working Paper.

MAZAR, N., O. AMIR, AND D. ARIELY (2008): "More Ways to Cheat—Expanding the Scope of Dishonesty," *Journal of Marketing Research*, 45, 651–653.

MILLER, N., P. RESNICK, AND R. ZECKHAUSER (2005): "Eliciting Informative Feedback: The Peer-Prediction Method," *Management Science*, 51, 1359–1373.

MILLER, N. H., J. W. PRATT, R. J. ZECKHAUSER, AND S. JOHNSON (2007): "Mechanism Design with Multidimensional, Continuous Types and Interdependent Valuations," *Journal of Economic Theory*, 136, 476–496.

MORRIS, S. AND H. S. SHIN (1998): "Unique Equilibrium in a Model of Self-Fulfilling Currency Attacks," *American Economic Review*, 587–597.

PALFREY, T. R. (2002): "Implementation Theory," in *Handbook of Game Theory with Economic Applications*, Elsevier, vol. 3, 2271–2326.

POSTLEWAITE, A. AND X. VIVES (1987): "Bank Runs as an Equilibrium Phenomenon," *Journal of Political Economy*, 95, 485–491.

PRELEC, D. (2004): "A Bayesian Truth Serum for Subjective Data," *Science*, 306, 462–466.

RUBINSTEIN, A. (1989): "The Electronic Mail Game: Strategic Behavior under 'Almost Common Knowledge'," *American Economic Review*, 385–391.

SALANIÉ, B. (2005): *The Economics of Contracts: A Primer*, The MIT Press.

Appendix

A Proof of Theorem 2

Clearly, s^{**} is a BNE; thus, it suffices to show the uniqueness. Suppose that s is a BNE. Fix an arbitrary $\omega \in \Omega$. Let

$$\alpha \equiv \min_{(i,t_i),\eta_i(t_i)=0} s_i(\omega,t_i)(\omega),$$

and

$$\tilde{T}_i \equiv \{t_i \in T_i \mid s_i(\omega,t_i)(\omega) = \alpha\} \text{ for each } i \in N.$$

Suppose that $\times_{i\in N} \bar{V}_i^{\infty}(E^*) = \emptyset$, that is, $\bar{V}_i^{\infty}(E^*) = \emptyset$ for all $i \in N$. Toward a contradiction, suppose that $\alpha < 1$, i.e., there exists an informed type who announces a dishonest message. It follows from (4) and (5) that $\tilde{T}_i \subset E_i^* \setminus T_i^{\dagger}$.

Consider an arbitrary $i \in N$ and $t_i \in T_i$ such that $s_i(\omega,t_i)(\omega) = \alpha$. By construction, $t_i \in \tilde{T}_i$ must be the case. From (4) and (5), α is equal to the average of the other agents' announcements on ω in expectation but not greater than any announcement. Hence, agent i expects that every other informed agent $j \neq i$ announces $m_j(\omega) = \alpha$, that is,

$$\pi_i\left(\times_{j\in N}\left(\tilde{T}_j \cup T_j^{\dagger}\right)\middle| t_i\right) = 1.$$

Accordingly,

$$\pi_i\left(E^* | t_i\right) = 1.$$

Hence, we have

$$\tilde{T}_i \subset \bar{V}_i^1(E^*).$$

Moreover, we have

$$\pi_i\left(\times_{j\in N}\left(\bar{V}_j^1(E^*) \cup T_j^{\dagger}\right)\middle| t_i\right) \geq \pi_i\left(\times_{j\in N}\left(\tilde{T}_j \cup T_j^{\dagger}\right)\middle| t_i\right) = 1.$$

Therefore,

$$\tilde{T}_i \subset \bar{V}_i^2(E^*).$$

Similarly, we have

$$\tilde{T}_i \subset \bar{V}_i^k(E^*) \text{ for all } k \geq 3.$$

Hence, we have

$$\tilde{T}_i \subset \bar{V}_i^\infty(E^*),$$

which, however, contradicts the supposition that $\bar{V}_i^\infty(E^*) = \emptyset$. Hence, we conclude $\alpha = 1$, that is, $s = s^{**}$, and thus, we have proved the "if" part.

Fix an arbitrary $\omega' \neq \omega$. We specify a strategy profile s^+ as follows: for every $i \in N$ and $t_i \in T_i$, whenever $\eta_i(t_i) = 1$, i.e., $t_i \notin T_i^\dagger$, we have

$$\begin{aligned} s_i^+(\omega, t_i) &= \omega \quad \text{if } t_i \notin \bar{V}_i^\infty(E^*), \\ s_i^+(\omega, t_i) &= \omega' \quad \text{if } t_i \in \bar{V}_i^\infty(E^*), \end{aligned}$$

and

$$s_i^+(\tilde{\omega}, t_i) = s_i^*(\tilde{\omega}, t_i) \quad \text{for all } \tilde{\omega} \neq \omega.$$

Clearly, from the previous argument, s^+ is a BNE, and $s^+ \neq s^*$ whenever $\bar{V}_i^\infty(E^*) \neq \emptyset$ for some $i \in N$. Hence, we have proven the "only-if" part.